SPEECH RECOGNITION OF NON-NATIVE SPEECH USING NATIVE AND NON-NATIVE ACOUSTIC MODELS

David A. van Leeuwen and Rosemary Orr

vanLeeuwen@tm.tno.nl
TNO Human Factors Research Institute.
Postbus 23,
3769 ZG Soesterberg,
The Netherlands.

R.Orr@kno.azn.nl
University Hospital Nijmegen St Radboud
Philips van Leydenlaan 15
5600 HB Nijmegen
The Netherlands

ABSTRACT

A speech recognition system is subjected to the speech of non-native speakers, using both native and non-native acoustic phone models. The problems involved with the mapping of phoneset from the non-native to native language are investigated, and a detailed analysis of phone confusions is made. For Dutch speakers, British English acoustic models give the best word recognition results.

INTRODUCTION

The study of speech as uttered by non-native speakers of the language has been a subject of research in phonetics for along time [1]. With maturing speech technology, the subject of non-native speech is becoming a topic of interest. Non-native speakers will form a new challenge for any technology for which acoustic training is an important factor, *e.g.*, codebook based coding systems, or speech and language recognition systems. One of the problems of training for non-native speakers is that the target group is very inhomogeneous—there are in principle as many potential non-native classes as there are languages in the world. This is a larger number than the number of dialects within a language, which has been the classic acoustic modelling challenge.

The standard approach for a technology such as speech recognition is to gather a database of the target group of users, and (re)train the system using this new database. For non-native speech, this means that if there are N languages for which speech recognition is available, a full matrix of native and non-native recognition systems would require N^2 speech databases, most of which will be non-native databases. Currently, the number of available non-native databases is very limited.

An alternative approach to non-native speech is to assume that non-native speakers will dominantly use their native phones, presumably by mapping the phones of the language they are speaking (L2) to their native language (L1). If this is the case—and the fact that foreign speakers can very well be characterized (and caricatured) supports this assumption—a speech recognition system can use the L1 phone models for the non-native speakers, combined with L2 dictionary and language models. In this way, only N acoustic training databases must be available for a full set of native and non-native recognition systems in N languages. Of course, there are non-native issues in pronunciation rules (dictionary) and language modelling as well, but we will not address these in this study.

This paper reports on an experiment for Dutch speakers speaking English, where a speech recognition system is trained with either native Dutch, British English or American English speakers. The main objective is to investigate whether L1 speakers should be recognized using L1 or L2 acoustic models when they are speaking in a non-native language L2. The implementation is limited—only one non-native language combination is investigated, using only one speech recognition system—and therefore the methodology of the experiment might have more implications than the bare results.

THE MIST SPEECH DATABASE

In late 1996 TNO recorded a speech database for Dutch continuous speech recognition named NRC0, similar to the Wall Street Journal corpus, (WSJ0) [2]. The main purpose of the database was to bootstrap the development of large vocabulary continuous speech recognition for the Dutch language. This database consisted of 52 speakers, each uttering 65 unique sentences. The sentence texts were taken from a Dutch newspaper (*NRC/Handelsblad*), read from a CRT screen in a quiet and low reverberant room, using a Sennheiser HMD 414-6 microphone, and high quality digital recording equipment. The number of speakers is smaller than for WSJ0 (and similar databases such as WSJCAM0 [3] and BREF80 [4]), and therefore TNO decided to extend the database in 1998 with another 80 speakers (NRC1). For these sessions, special sentences were recorded additional to the 65 utterances for continuous speech recognition systems. These included 'foreign language sentences,' which were sentences in English, French and German. The prompt texts for the foreign language sentences were taken from newspaper texts, English from *Wall Street Journal*, German from *Frankfurter Rundschau* and French from *Le Monde*. These were the same sources from which the development and

test sentences in the SQALE project [5] were chosen. The recordings of the foreign language sentences can be considered non-native speech material for English, French, and German.

The majority of the speakers for NRC1 were recruited from the institute. Of the institute's employees, 60% has an academic background, and 20% a higher technical education. This is not a representative sample of the Dutch population. There is the advantage, however, that a relatively high fraction of subjects can be expected to be able to speak one or more foreign languages. It was left to the subject's own discretion to decide whether or not to record the foreign sentences. Thus, of the 74 subjects that recorded foreign speech, 71 recorded English, 66 German and 60 French. The prompt texts consisted of five sentences that were the same for all speakers, and could function as adaptation sentences. A further five sentences were chosen, which were unique for every speaker.†

For the purpose of the MIST workshop, TNO decided to share the non-native speech data with other research institutions. A liberal license agreement allows people to use the speech material for research purposes, free of charge. As a reference, 10 Dutch sentences per speaker were added to the non-native speech database, again consisting of 5 sentences that were the same across all speakers, and 5 unique sentences. Thus a total of over 5 hours of speech is available for the scientific community. Only for the Dutch sentences, a detailed orthographic transcription could be made, for the other three languages just the prompt texts were distributed. It is hoped that native speakers at other institutes will provide the community with corrected transcriptions.‡ A number of articles in these proceedings [6, 7] already report on results using this database. For the experiments in this paper, only the English utterances were used.

THE ABBOT SPEECH RECOGNIZER

For the speech recognition system used in this experiment, we used the Abbot large vocabulary continuous speech recognition system [8]. Abbot is a hybrid neural net/Markov model recognition system. The most important difference from traditional hidden Markov model systems is that the neural net directly estimates *a posteriori* phone probabilities for each speech frame. The forward pass in the recurrent neural net can be calculated quickly, and phone probabilities are quite well determined. This makes the decoding search relatively easy, and therefore the system is known for its fast recognition speed. By choosing the appropriate decoder, both a phone recognition system and a word recognition system can be built.

The components needed for the various word recognizers are

- L2 (English) and L1 (Dutch) acoustic models
- L2 dictionary
- L2 to L1 phoneset mapping
- L2 language model.

When Dutch acoustic models are used, a dictionary of English words in terms of Dutch phones is needed. One way to achieve this is to use an English dictionary, and to translate all English phones into corresponding Dutch phones. For this the reason the L2 to L1 phoneset mapping is necessary.

For a phone recognizer, the phone mappings appear to be unnecessary. However, for evaluation of the Phone Error Rate (PER) a phone level reference transcription is needed. Because the test database is annotated at the word level, a dictionary is needed to convert the L2 reference words into L1 phones. As English dictionaries in terms of Dutch phone sets are not available, the phone mapping is necessary in this case as well.

EXPERIMENTAL SETUP

The test database used is the English part of the MIST speech database. The speakers were separated into two groups, training and testing speakers. The training speakers were not used in this experiment, and only the five unique sentences per speaker were used. This resulted in 180 utterances by 36 speakers. Of the 3147 words 129 (4%) words were Out Of Vocabulary (see below).

Table I. Acoustical training conditions for three languages.

Language	American	British	Dutch
Database	WSJ0	WSJCAM0	NRC0
# speakers	84	90	48
speech length (hr)	13	13	7
phones	53	44	39
phoneset	ICSI/LIMSI	BEEP	CELEX

Three different acoustical models were used, American English, British English, and Dutch. The training conditions are comparable, except for Dutch, which has about half the training time (7 hr). The Abbot speech recognition system is known to have a relatively quickly saturating performance with increasing training data, due to the limited number of parameters to be estimated. In table I the acoustical conditions are tabulated. The phoneset for Dutch is a subset of the phoneset defined in the CELEX dictionary [9]. For American English, the ICSI/LIMSI phoneset is used [10, 11]. The training for American and British English was performed by Cambridge

† For each language, there are 2–5 sentences that occur twice among the speakers, due to an unfortunate misconfiguration during the sentence selection.

‡ The latest transcriptions can always be found at URL `ftp://ftp.tm.tno.nl/pub/speech/mist`.

Table II. The phone map used in order to translate the American and British English dictionaries using the Dutch phoneset. The second and fourth conlumn show the full English phonesets, the middle column shows the Dutch phones to which the phones are mapped. The phones f, h, ʤ, l, m, n, ŋ, s, ʃ, v, j, z, ʒ occur in all three phone sets, and are not shown.

American→		Dutch	←British	
bottle	ɑ	ɑ	ɑ	heart
hamm	æ	ɛ	æ	zap
sum	ʌ	ɑ	ʌ	rough
might	aɪ	a: j	aɪ	ice
more	ɔ	ɔ	ɔ	lord
		ɔ	ɒ	pot
ago	oʊ	o:	əʊ	rogue
annoyed	ɔɪ	ɔ j	ɔɪ	boil
house	aʊ	aʊ	aʊ	house
again	ə	ə	ə	again
alive	l̩	ə l		
atom	m̩	ə m		
heaven	n̩	ə n		
after	ɹ̩	ə ʀ		
hurd	ɝ	ə ʀ	ɜ:	burn
bet	ɛ	ɛ	ɛ	bet
		ɛ ə	ɛə	hair
pain	eɪ	e:	eɪ	pain
adding	i̯	ɪ		
fit	ɪ	ɪ	ɪ	fit
		ɪ ə	ɪə	here
beat	i	i:	i	beat
hook	ʊ	u:	ʊ	bush
cool	u	u:	u	cool
		u: ə	ʊə	poor
lobe	bˀ	b	b	board
bow	b			
beach	ʧ	t ʃ	ʧ	beach
shed	dˀ	d	d	does
does	d			
this	ð	d	ð	that
butter	ɾ	d		
jig	gˀ	g	g	go
go	g			
aha	ɦ	h		
arc	kˀ	k	k	cow
cow	k			
chip	pˀ	p	p	pot
pot	p			
raise	ɹ	ʀ	ɹ	raise
fit	tˀ	t	t	tip
tip	t			
thing	θ	t	θ	thing
walk	w	ʋ	w	walk

University. In the American English training procedure, a different dictionary was used [12].

The size of the vocabulary was conservatively chosen to be 20k words. The limited size was used because it was not an objective to optimize a system for performance, but rather to compare performances. The vocabulary and dictionaries were effectively determined by the freely available demonstration version of Abbot [13]. The American English pronunciation dictionary is based on the CMU dictionary [14], whereby the phoneset was converted using an automatic phone mapping to the ICSI phone set. The British English dictionary is a subset of the BEEP dictionary [15]. In order to obtain dictionaries for the Dutch phone set, both dictionaries were translated using a phone map shown in table II.

The language model used is a 20k word trigram language model, which was developed using American English texts pre-dating spring 1998. The decoder used for Abbot is 'chronos,' a time-synchronous stack decoder [16]. The language model was used for all word recognition runs, except for the Dutch baseline run.

Phone mapping

The phone mapping shown in table II needs some explanation. It was based on our phonetic intuition of the similarity between Dutch and English phones. The table shows only one mapping per phone, but later we will show that experiments have been carried out with multiple mappings. Some phone mappings have been made consistent with the way the Dutch vocabulary, that was used in the acoustic model training, expresses words in terms of the Dutch phones. For instance, the mapping [aɪ] → [a: j] is chosen over [a: ɪ], because the CELEX dictionary has entries for words like *haai* → [h a: j] (shark). In the training process of the Dutch acoustic models, therefore, the [j] models the [ɪ] in the context of [aɪ].

The American English phoneset has separate entries for 'closures,' plosives without an audible release, [bˀ, dˀ, gˀ, kˀ, pˀ, tˀ], in combination with the standard IPA plosives. In the dictionary used for American English, most occurrences of a plosive are preceded by a closure, *e.g.*, *bee* → [bˀbi]. However, the non-audible release can stand on its own, *e.g.*, as in *add* → [ad]. For this reason, the closures are mapped to Dutch plosives, and the plosives are mapped to nothing.

BASELINE RESULTS

In order to have a reference for the experiments with non-native speech, a number of baseline tests were performed. For this, development test material used in the SQALE project was used. This consisted of 20 native speakers for American and British English (10 male, 10 female). For a baseline for the Dutch models, 20 speakers of the NRC1 Dutch database were used. Each of the speakers contributes 10 utterances to the test. In table III the phone and word errors of the recognizer are given. The baseline results are only indicative of the recognition system; they are not 'optimal' values. For instance, the language model has not been optimized for the speech domain. In determining the PER for English, an automatic expansion of the reference word transcriptions has been made, using the appropriate dictionary. Because the English dictionaries have multiple pronunciations per word,

many arbitrary decisions have been made in generating the phone reference transcription. This leads to an estimation for the PER which is too high.

The Dutch dictionary has only single pronunciation entries. This may be the reason that the PER figure is lower than for English. The word error rate for Dutch is much higher than for English. The Dutch language model was based on a 78 million words text of newspaper text, defining the vocabulary as the most frequent 20 000 words. The language model was built specifically for the baseline test, and has not been optimized.

Table III. Word and phone error rates (WER) in % for baseline conditions. The top line gives the WER in the standard 'forward' condition, that is used throughout this work. 'Forward' means a forward pass only, 'fw/bw' means forward and backward pass (see text).

Language	American	English	Dutch
WER (forward)	**27.6**	**26.2**	**37.7**
WER (fw/bw)	22.4	22.5	34.4
PER (forward)	39.8	37.4	35.6
PER (fw/bw)	37.3	34.7	33.4

In table III results for a forward/backward pass are given as an indication as to how much lower the error rates are if the posterior log probabilities are averaged with 'backward' runs. Because Abbot utilizes a recurrent neural network, past acoustic context is automatically modelled. In order to model future acoustic context, a 'backwards' network can be trained by feeding the network acoustic features that are reversed in time. In the recognition pass, the backwards classified phone probabilities can be merged with the forward stream, which generally leads to lower error rates.

Results for non-native speech

In table IV the results for the MIST database are given. Results obtained with Dutch models are made with either US or UK dictionary, translated using the phone map of table II.

Table IV. The word and phone error rates (in %) for the MIST database of Dutch speakers speaking English. A 20k English vocabulary and an accompanying trigram language model was used. (US, UK, NL) means American, British, Dutch. The standard deviation of the numbers is approximately 0.8 %.

Acoustic models	US	UK	NL	NL
Pronunciation dictionary	US	UK	US	UK
WER	68.8	60.9	68.9	73.4
PER	55.7	49.1	54.5	56.2

It appears that British acoustic phone models give the lowest error rates for the Dutch MIST speakers, both in phone and word error rate. It is interesting to note, that the difference between PER and WER is smaller—and has actually reversed sign—with respect to the baseline. One is tempted to assign this to a language model incompatibility, but this is unlikely because of the very similar source of both tests, namely the SQALE sentences.

Influence of the phone mapping

Of the results of table IV, the last two columns are most interesting, because they involve a non standard combination of L1 acoustic models and L2 language models. The phone mapping shown in Table II is the first mapping we tried, based on phonetic intuition. The Dutch phoneset contains some phones that are not covered by the English mappings, namely [ø:, ei, ʉ, œy, x], and [y:]. Other English phones, that experienced speakers are capable of using, have no real Dutch equivalent. Examples of these are the infamous 'th' consonants [θ] and [ð]. We experimented with a couple of changes to the phone mapping, in order to investigate if any of them would lower the word error rate.

First, we adapted the dictionary conversion tool to accept alternatives for phone conversions. This involved a recursive expansion of alternative pronunciation strings. For instance, if both the alternatives [ð] → [d|z] and [ʌ] → [ɑ|ʉ] are allowed, the word 'mother' ([mʌðɹ]) gets four alternative pronunciations, [mɑdəʀ, mʉdəʀ, mɑzəʀ] and [mʉzəʀ]. The inclusion of the above examples and [θ] → [t|s] lead to an *increase* in word error rate of 7 %-point for the American English dictionary. Apparently, allowing more pronunciation variants per word causes more options for erroneous words than that it helps to find options for the correct word.

We have run several tests in order to investigate what the individual contribution of the alternatives to this increase is. The alternatives that we defined for American and British pronunciation are shown in the first columns of table V, together with the difference in WER the individual alternative makes. Again, almost all alternatives lead to an *increase* in word error rate.

Table V. Changes from the default phone mapping (see table II). In the last column, the increase in the word error rate (in %-point) with respect to the baseline is given.

English	Dutch	US	UK
ʌ	ɑ|ʉ	+2.6	+1.4
ʌ	ʉ	+4.3	+3.0
θ	t|s	+0.5	+2.1
θ	s	+0.8	+2.1
ð	d|z	+2.9	+1.8
ɾ	d|t	−0.8	
ɾ	t	0.0	
aɪ	a: i:	+2.2	+3.4
aɪ	a: ɪ	+2.0	+2.3
i	ə	+0.8	
ɛə	ɛ ʀ		−0.2
ʊə	u: ʀ		−0.5

The increase of PER for the mapping [ʌ] → [ʉ] surprised us, because in the stereotypical Dutch En-

Table VI. Individual phone confusions. Only phones that are confused more often with others (left number) than that they are recognized correctly (right number) are shown. The leftmost columns show the phone confusion considered. The second three columns show the phone confusion numbers for the baseline tests. The third and fourth three columns show the confusion numbers for the non-native database. Boldface indicates more errors than correct. In the case of the Dutch phoneset (last group of rows), a dictionary phone mapping for the reference transcription was used.

Phones			Reference			Non native database MIST					
Set	ref.	rec.	test	err	corr	mapping	err	corr	mapping	err	corr
US	m̩	m	SQALE	**27**	2	US	**29**	1			
	n̩	n		**163**	42		**332**	56			
	ɦ̩	h		**2**	1		**5**	1			
	ʒ	ʃ		0	6		3	3			
	i̩	ɪ		**146**	74		**207**	53			
UK	a	ɛ	SQALE	9	126	UK	**112**	55			
	ʒ	ʃ		0	6		**2**	1			
NL	ɑ	aː	NRC1	141	2679	UK	**184**	166	US	74	183
	ɔ	oː		254	1580		223	340		**181**	167
	ʊ	r		3	225		**36**	7		0	0
	ʒ	ʃ		**15**	0		**2**	0		**3**	0
	ʤ	j		**6**	0		**18**	0		**19**	0
	f	v		218	517		**175**	161		**178**	161
	ɡ	x		4	4		**18**	4		**9**	5
	ɡ	k		**11**	4		**80**	4		**81**	5
	v	f		303	1968		**216**	164		**214**	162
	z	s		214	1221		**268**	149		151	266

glish accent [ʌ] is pronounced as [ʉ]. The reason might be, that the acoustic modelling for [ʉ] in Dutch is relatively poor. The confusibility of [ʉ] with [ə] is high because the schwa lies acoustically very close to the unstressed [ʉ].

One more elaborate expansion is that of the plosives in the American English phone set. The mapping of [b, d, g, k, p, t] to nothing leads to a few errors in the converted dictionary. Words like *update* have the US expansion [ʌpˈdeɪtˈ], where there is a closure of /p/ followed by the release /d/. In our original mapping, the latter phone was deleted. Correcting for these occurrences (translating *update* → [ɑpdeːt]) lead to a *decrease* of the word error rate for the American dictionary of 0.3 %-point. A combination of this with the alternative [ɾ] → [d|t] lead to a total decrease of 1.2 %-point.

Individual phone scores

By investigating the phone recognition result, it is possible to make an inventory of the individual phone scores. A phone class based alignment algorithm [17] can provide a fairly good measurement of the phone confusion matrix, even for continuous speech recognition. A way to summarize the problems in phone recognition is to tabulate the phones that are recognized more often as a different phone than as themselves. In table VI these phones are indicated for a number of baseline and non-native tests.

In some cases we can conclude that the basic models are not well trained. This is the case for, *e.g.*, the American [m̩, n̩] and [i̩], and the Dutch [ʒ, ʤ] and [ɡ]. But for other phones, there is a clear effect of the non-native speech. From table VI it is clear that the British [a] is pronounced closer to the [ɛ] by the Dutch speakers. When the Dutch phoneset is used for the non-native speaker, there are many examples of phones that have a high confusibility with others. This may be an artifact of the automatic dictionary mapping. Interestingly enough, both [f] and [v] have a tendency to be interchanged in recognition with respect to the dictionary expansion. Our understanding of this is that in Dutch local accents, the /f/ and /v/ have acoustic realizations that are similar, because the difference in voicing tends to blur.

CONCLUSIONS

We have shown a methodology that allows non-native (L2) speech recognition using native (L1) speech models, L2 dictionary and grammar, and an L2 → L1 phone mapping. In the case of Dutch non-native speakers of English, the plain word recognizer using British English models gives lower word error rates than the approach given above, but it is not known whether this will generalize to other combinations of non-native speech. Still, the word error rate of the non-native speakers is a factor 2 higher than for native speakers. The phone mapping, necessary in order to define a L2 dictionary in terms of L1 phones, forms a weak link in the approach. A more elaborated rule based translation of the vocabulary should lead to better results for the approach taken here.

REFERENCES

[1] James Emil Flege, Ocke-Schwen Bohn, and Sunyoung Jang. Effects of experience on non-native speakers' production and perception of english vowels. *Journal of Phonetics*, 25:437–470, 1997.

[2] D. B. Paul and J. M. Baker. The design for the Wall Street Journal-based CSR corpus. In *Proc. Fifth DARPA Speech and Natural Language Workshop*, pages 357–362. Morgan Kaufmann Publishers, Inc., 1992.

[3] Jeroen Fransen, David Pye, Tony Robinson, Phil Woodland, and Steve Young. WSJCAM0 corpus and recording description. CD-ROM documentation, 1994. CUED Cambridge (UK).

[4] L. F. Lamel, J. L. Gauvain, and M. Eskenazi. BREF, a large vocabulary spoken corpus for French. In *Proc. Eurospeech*, 1991.

[5] S. J. Young, M. Adda-Dekker, X. Aubert, C. Dugast, J.-L. Gauvain, D. J. Kershaw, L. Lamel, D. A. van Leeuwen, D. Pye, A. J. Robinson, H. J. M. Steeneken, and P. C. Woodland. Mutilingual large vocabulary speech recognition: the European SQALE project. *Computer Speech and Language*, 11:73–89, 1997.

[6] R. Wanneroy, E. Bilinski, C. Barras, M. Adda-Decker, and E. Geoffrois. Acousic-phonetic modeling of speech for language identification. In *These Proceedings*, pages 9–13, 1999.

[7] Geoffrey Durou. Multilingual text-independent speaker identification. In *These Proceedings*, pages 115–118, 1999.

[8] Tony Robinson, Mike Hochberg, and Steve Renals. *The use of recurrent networks in continuous speech recognition*, chapter 7, pages 233–258. Kluwer Academic Publishers, 1996.

[9] R. H. Baayen, R. Piepenbrock, and L. Gulikers. The CELEX lexical database. CDROM, Linguistic Data Consortium, University of Pennsylvania, Philadelphia, PA, 1995.

[10] J. L. Gauvain, L. F. Lamel, G. Adda, and M. Adda-Decker. Speaker-independent continuous speech dictation. *Speech Comm.*, 15:21–37, 1994.

[11] D. J. Kershaw. *Phonetic Context-Dependency in a Hybrid ANN/HMM Speech Recognition System.* PhD thesis, University of Cambridge, January 1997. URL: `http://www-svr.eng.cam.ac.uk/~djk/Publications/thesis.html`.

[12] Tony Robinson. Private Communication.

[13] See URL. `http://svr-www.eng.cam.ac.uk/~ajr/abbot.html`.

[14] See URL. `http://www.speech.cs.cmu.edu/cgi-bin/cmudict`.

[15] See URL. `ftp://svr-ftp.eng.cam.ac.uk/pub/comp.speech/dictionaries`.

[16] Tony Robinson and James Christie. Time-first search for large vocabulary speech recognition. In *ICASSP*, 1998.

[17] David A. van Leeuwen and Michael de Louwere. Objective and subjective evaluation of the acoustic models of a continuous speech recognition system. In *Proc. Eurospeech*, pages 1915–1918, 1999.